# AKSHAYARAM SWAMINATHAN

Senior GenAI / AI-ML Engineer | Production LLM Systems · RAG · Agentic AI · MLOps
akshayram.swami@gmail.com | linkedin.com/in/akshay-swami

## PROFESSIONAL SUMMARY

Generative AI (GenAI) and Machine Learning (ML) Engineer with 3+ years designing, building, and shipping production-grade Large Language Model (LLM) systems end to end. Sole founding engineer on multiple products built from scratch on Amazon Web Services (AWS) and Google Cloud Platform (GCP) — including a Natural Language-to-SQL (NLP-to-SQL) query engine that cut per-query inference cost ~60%, an agentic GraphRAG assistant over enterprise systems, and a multi-tenant Software-as-a-Service (SaaS) chatbot platform. Deep expertise in LLM fine-tuning (LoRA, QLoRA, PEFT), Retrieval-Augmented Generation (RAG), multi-agent orchestration, and MLOps. Track record of leading small engineering teams and turning research and proofs of concept (PoCs) into secure, scalable, observable systems.

## TECHNICAL SKILLS

**Generative AI & LLMs:** LLM fine-tuning (LoRA, QLoRA, Parameter-Efficient Fine-Tuning / PEFT), Retrieval-Augmented Generation (RAG), GraphRAG, NLP-to-SQL, prompt engineering, Chain-of-Thought, ReAct, multi-agent orchestration, Reinforcement Learning from Human Feedback (RLHF)

**Frameworks & Libraries:** LangChain, LangGraph, Hugging Face Transformers, PyTorch, TensorFlow, Sentence Transformers, OpenAI SDK, Google GenAI SDK, FastAPI, Playwright

**Cloud — AWS:** Amazon Bedrock (Claude, Titan), SageMaker, Lambda, API Gateway (REST + WebSockets), EC2, S3, RDS, DynamoDB, ElastiCache (Redis), Kinesis, Amazon Lex, Amazon Connect, CloudFront, QuickSight, AWS CDK

**Cloud — GCP:** Vertex AI (Pipelines, Model Registry, Evaluation, Agent Builder), AutoML, Gemini, BigQuery, Cloud Run, Cloud Functions, Pub/Sub, Cloud SQL, Firestore, Looker Studio

**MLOps & Data:** MLOps, CI/CD for ML, Vertex AI Pipelines, Extract-Transform-Load (ETL/ELT), data lake architecture, AWS Glue, Integrated Workplace Management System (IWMS) integrations (IBM TRIRIGA, Maximo)

**Vector & Databases:** Qdrant, Pinecone, PGVector, Amazon OpenSearch, Neo4j, Redis, PostgreSQL, BigQuery

**Programming:** Python (primary), SQL, R

## PROFESSIONAL EXPERIENCE

**Quantum Strides LLC** — AI & Automation Engineer — Glen Allen, VA, USA (Remote) — Jul 2025 – Present

- Sole founding engineer: designed, built, and shipped the flagship platform's Release 1 end to end, then scaled engineering by recruiting and leading a 5-developer team to extend the platform.
- Delivered two production AI products end to end on AWS serverless — an NLP-to-SQL engine spanning 300+ tables across 20+ facility schemas, and a multi-tenant white-label SaaS chatbot — owning architecture, Role-Based Access Control (RBAC), audit logging, and licensing.
- Reduced per-query Amazon Bedrock inference cost ~60% through Redis semantic caching and tiered model routing (Claude Haiku for lookups, Claude Sonnet 4 for complex analytical SQL) with dry-run validation.
- Achieved zero-downtime production releases through AWS CDK blue/green deployments, with tenant data isolation, Personally Identifiable Information (PII) redaction, and per-tenant rate limiting enforced across products.

**Rapyder Cloud Solutions** — AI/ML Engineer → Senior GenAI/ML Engineer — Bangalore, IN — May 2023 – Jul 2025

- Led the GenAI technical practice — a 7-engineer team delivering enterprise Generative AI across Fintech, Healthcare, Retail, and GovTech; mentored 6 junior engineers and ran cross-functional architecture reviews.
- Designed and delivered 12+ production GenAI systems, driving 17% average operational-efficiency gains across client verticals through LLM-powered automation.
- Fine-tuned GPT-based models for domain-specific Natural Language Processing (NLP) tasks using PEFT/LoRA, improving generation quality 60% over zero-shot baselines on client benchmarks.
- Built AWS SageMaker MLOps pipelines (automated training, evaluation, versioning, CI/CD deployment), cutting release-cycle time 30%.
- Engineered a low-latency audio + text streaming pipeline (AWS Kinesis, Lambda, API Gateway WebSockets) serving 500+ concurrent users at 50% lower end-to-end processing time.
- Reduced customer-support ticket volume 40% with an Amazon Lex + Amazon Connect AI call agent (intent detection, slot filling, authentication, CRM ticketing); contributed to Rapyder's AWS Generative AI Competency — the 3rd company in India to earn it.

**NimbleWork Inc (formerly Digité)** — Generative AI Intern — Bangalore, IN — Feb 2023 – May 2023

- Integrated GPT-4 and open-source LLMs into kAIron, a microservice-based chatbot-training platform — building intent classification, entity extraction, and context-aware response generation.
- Built a ClickStream analytics dashboard (Python + Plotly) tracking cross-property session flows; surfaced the top-3 friction points, contributing to a 25% uplift in measured user satisfaction.

## KEY PROJECTS

**NL-to-SQL Facility Intelligence Platform (product name under employer IP) | Quantum Strides · Production · AWS Serverless**

*Stack: Python, AWS Lambda, Amazon Bedrock (Claude Sonnet 4), RDS PostgreSQL, ElastiCache (Redis), API Gateway, S3, AWS CDK*

- Built end to end on AWS serverless: query → schema retrieval from S3 → dynamic few-shot prompt assembly → Claude Sonnet 4 (Bedrock) SQL generation → RDS PostgreSQL execution — covering multi-join, aggregation, and time-series queries across 20+ facility schemas and 300+ tables.
- Cut per-query Bedrock inference cost ~60% with a Redis + sentence-transformer semantic cache that short-circuits semantically similar queries.
- Added tiered model routing (Claude Haiku for lookups, Claude Sonnet 4 for complex analytical SQL) with SQL-syntax validation and execution dry-runs as hallucination guardrails.
- Built Lambda-based ETL ingesting heterogeneous IWMS sources (IBM TRIRIGA, Maximo) into an RDS PostgreSQL data lake with schema normalization, delta detection, and data-quality checks; full RBAC, audit logging, and licensing via AWS CDK.

**AI Knowledge & Walkthrough Assistant for an Enterprise IWMS — Agentic GraphRAG | Quantum Strides · Pilot · Agentic AI**

*Stack: Python, FastAPI, LangGraph, Amazon Bedrock (Claude), Neo4j, Pinecone, Playwright / browser-use, React*

- Built an agentic ingestion pipeline that autonomously learns an enterprise TRIRIGA instance: Playwright + browser-use drive the live UI while Claude (Bedrock) interprets it, producing structured knowledge artifacts (page, form-schema, workflow models) stored in Neo4j (graph) + Pinecone (vector) for GraphRAG retrieval.
- Engineered a code-enforced never-mutate safety guard (not prompt-based) — a guarded browser-tool layer that physically blocks any Submit/Save/Delete action, verified in supervised runs against a production-grade instance — zero mutations, ever.
- Solved TRIRIGA's single-session constraint with a single-producer/multi-consumer concurrency model that serializes browser access while parallelizing LLM calls, plus session re-authentication and retry so multi-hour runs survive timeouts.
- Wired a LangGraph chat pipeline (intent classification → GraphRAG retrieval → synthesis) with a guarded live-walkthrough mode driving the real UI; cut per-form extraction ~2.7× (≈20→7 min), added per-run Bedrock cost instrumentation, and shipped under strict Test-Driven Development (TDD) with 900+ passing tests.

**Multi-Tenant White-Label SaaS Chatbot (product name under employer IP) | Quantum Strides · Production · AWS Serverless**

*Stack: Python, AWS Lambda, Amazon Bedrock (Claude Sonnet 4), API Gateway WebSockets, Qdrant (ECS Fargate), DynamoDB, S3, CloudFront, AWS CDK*

- Built a serverless multi-agent pipeline: API Gateway WebSocket → Router Lambda → per-tenant Qdrant knowledge base (ECS Fargate) → Claude Sonnet 4 (Bedrock) synthesis, with conversation state persisted in DynamoDB.
- Shipped an embeddable JavaScript widget via CloudFront with self-serve tenant onboarding, admin portal, connector setup (REST, webhooks, Confluence, Notion), and usage analytics.
- Enforced tenant isolation via separate Qdrant namespaces, query-time access controls, pre-context PII redaction, and per-tenant API Gateway rate limiting.

**Streaming Multilingual Voice Agent (PoC) | GCP · Vertex AI**

*Stack: GCP Speech-to-Text v2 (Chirp), Vertex AI Gemini 1.5 Flash, Cloud Text-to-Speech (Journey), Cloud Run, Pub/Sub, Firestore*

- Built a GCP-native streaming voice agent (speech-to-text → LLM → text-to-speech) with automatic language detection across 10+ languages and sub-1.5s perceived latency by generating responses on interim transcripts; Cloud Run autoscaled sessions with per-session state in Firestore.

**ProfSam — Ed-Tech Company Built Solo (Founder-Engineer) | profsam.com · Live**

*Stack: Next.js, Sanity CMS, Vertex AI (Veo 3.1), PostgreSQL, AWS Fargate, YOLOv8, PaddleOCR*

- Designed, built, and run my family's ed-tech company end to end, solo: marketing site + SEO, the ScoreLab CBSE learning platform, an NL-to-SQL admissions chatbot, an AI video studio (Vertex AI Veo 3.1 with keyless AWS ↔ GCP auth), college-data pipelines with OCR, and all cloud infrastructure.
- 16k new users in 90 days from organic search (GA4).

## CERTIFICATIONS

AWS Certified Machine Learning – Specialty · AWS Certified Cloud Practitioner · BrainBench Certified Python Developer

## EDUCATION

BSc Data Science — CHRIST (Deemed to be University), Lavasa, Pune — 2020 – 2023
International Baccalaureate Diploma Programme (IBDP) — The Indian Public School, Chennai — 2018 – 2019

## EVENTS & ACHIEVEMENTS

Winner, 24-hour Hackathon – CHRIST University (2022) · AWS Generative AI Partner Hackathon (2024) · AWS Generative AI Workshop & Data/Database/Analytics Roadshow (2023)